

2023年度 秋学期

# 卒 業 論 文

## 大規模言語モデルを用いた ナレッジグラフ抽出手法

指導教員: 桑原 和宏

立命館大学 情報理工学部

卒業研究 3 (CB)

コース: 先端社会デザイン

学生証番号: 2600200381-2

氏名: 松見 直哉



# 概要

ナレッジグラフはセマンティックネットワークとも呼ばれ、実世界のエンティティとその相互関係をモデル化する技術である。この技術は、テキストをはじめとする非構造化データの構造化を意味し、情報検索や推論の精度向上に役立っている。しかし、実用的なナレッジグラフを構築するためには、辞書や機械学習に必要な学習データが必要であり、コストがかかる。これは、未だ構造化されていない知識の構造化を妨げる要因であり、本研究ではその解消に焦点をあてた。具体的には、大規模言語モデルを用いて、テキストデータからナレッジグラフを抽出する手法を提案する。実装にあたり、グラフ構造化知識の事前学習なし且つ自然言語処理を用いた従来手法を参考にした。従来手法で行われる各プロセスに適したプロンプトに置き換え、大規模言語モデルに逐次的に与えた。評価実験では、提案手法と従来手法で同様のテキストデータからナレッジグラフを抽出して比較した。結果、提案手法が従来手法に、情報量と正確性の両面で優位であることが示された。その要因は、テキストとナレッジグラフを細かく分析した結果、文脈把握能力と情報抽出精度の高さにあると考察する。



# 目次

図目次	vii
表目次	ix
リスト目次	xi
<b>第 1 章 はじめに</b>	<b>1</b>
1.1 研究の背景 . . . . .	1
1.2 研究の目的 . . . . .	2
1.3 論文の構成 . . . . .	2
<b>第 2 章 関連研究・基盤技術</b>	<b>3</b>
2.1 関連研究 . . . . .	3
2.2 基盤技術 . . . . .	3
<b>第 3 章 大規模言語モデルを用いたナレッジグラフ抽出手法</b>	<b>5</b>
3.1 概要 . . . . .	5
3.2 提案手法 . . . . .	5
<b>第 4 章 実験・評価・考察</b>	<b>13</b>
4.1 実験 . . . . .	13
4.2 評価 . . . . .	13
4.3 考察 . . . . .	14
<b>第 5 章 おわりに</b>	<b>27</b>
謝辞	29
参考文献	31

付録 A	ソースコード	33
付録 B	ナレッジグラフ	37

# 目次

4.1	従来手法 (テキスト 1)	15
4.2	提案手法 (テキスト 1)	15
4.3	従来手法 (テキスト 2)	16
4.4	提案手法 (テキスト 2)	16
4.5	従来手法 (テキスト 3)	18
4.6	提案手法 (テキスト 3)	18
4.7	従来手法 (テキスト 4)	20
4.8	提案手法 (テキスト 4)	20
4.9	従来手法 (テキスト 5)	21
4.10	提案手法 (テキスト 5)	21
4.11	従来手法 (テキスト 6)	22
4.12	提案手法 (テキスト 6)	22
4.13	従来手法 (テキスト 7)	23
4.14	提案手法 (テキスト 7)	23
B.1	従来手法 (テキスト 8)	37
B.2	提案手法 (テキスト 8)	37
B.3	従来手法 (テキスト 9)	38
B.4	提案手法 (テキスト 9)	38
B.5	従来手法 (テキスト 10)	39
B.6	提案手法 (テキスト 10)	39





# 表目次

3.1	エンティティの抽出 . . . . .	7
3.2	エンティティの属性分類 . . . . .	8
3.3	リレーションの抽出 . . . . .	10
3.4	リレーションの抽出 (続き) . . . . .	11
3.5	RDF トリプルの記述 . . . . .	12
4.1	ノード数 . . . . .	14
4.2	正答数/エッジ数 . . . . .	14



# リスト目次

3.1 プログラム準備段階 . . . . .	6
A.1 ソースコード . . . . .	33



# 第 1 章

## はじめに

### 1.1 研究の背景

ナレッジグラフはセマンティックネットワークとも呼ばれ、実世界のエンティティとその相互関係をモデル化する技術である。テキストデータなどの非構造化データを構造化できるため、検索エンジンや質問応答システムなどに活用され、情報検索や推論の精度向上に役立っている。人間のもつ知識のほとんどは文章や会話をはじめとする自然言語で表現され構造化されていないため、ナレッジグラフを用いて構造化することが有効なのである。

しかし、テキストデータからナレッジグラフを作成することは簡単に実現できるものではない。テキストからエンティティとリレーションを抽出する手法は手作業のほか、あらかじめ辞書を用意する手法 [1] と、機械学習を活用した自然言語処理手法 [2] がある。辞書や機械学習では、複雑な計算が要求されたり、大規模な学習データを必要としたりし、コストがかかる [3, 4]。よって、いずれの手法を用いても、実用的なナレッジグラフを素早く簡単に実現することは難しい。

辞書や機械学習を用いず、低コストでナレッジグラフが作成できるようになることは、人間の知識のうち大部分を占める構造化されていない知識の多くを、リアルタイムにモデル化できるようになることを意味する。そして、情報検索や推論の精度の飛躍的な向上に直結し、情報サービスがより豊かに展開される。例えば、大規模言語モデルは、事実知識の把握という点で完璧には程遠い [5] が、ナレッジグラフを用いて知識を補填することでより高いパフォーマンスを発揮する [6]。従って、実用的なナレッジグラフを低コストで構築する手法を実現する価値は大いにある。

## 1.2 研究の目的

本研究の目的は、実用的なナレッジグラフを、辞書や機械学習を用いず、それらよりも容易な方法で構築する手法の提案である。そこで、大規模言語モデルに着目した。2023年11月に ChatGPT が公開されて以来、その自然言語の理解や生成能力の高さは周知のものとなっている。従来の自然言語処理モデルを用いて行われていたタスクは、大規模言語モデルを用いることで、より簡単により良い結果を得られる可能性がある。ナレッジグラフの構築もそのひとつと言え、大規模言語モデルを活用することにした。

また、今回ナレッジグラフを抽出するテキストデータとして、ニュース記事を用いた。その理由は2点ある。まず、新聞社による記事は、文章校正やファクトチェックがなされており、自然言語の用法とその内容の正当性に信頼を置くことができるという点である。つぎに、ニュース記事はその性質上、テキスト量に対する情報量が多く、実装の変化をナレッジグラフに、より反映するという点である。

以上により、大規模言語モデルを活用して、テキストデータであるニュース記事からナレッジグラフを抽出する。その手順は従来の自然言語処理を用いた手法をなぞる。大規模言語モデルは高い能力をもつが、テキストからナレッジグラフを抽出してくださいと命令しても、期待されるものは得られない。従って、命令を細かく分けたプロンプトエンジニアリングを行うため、テキストがナレッジグラフになる一連のプロセスを紐解き応用した。

完成した提案手法は、その情報量と正確性を、従来手法と比較して評価する。提案手法が優位であることが示されれば、本研究の目的は達成されたことになり、未だ構造化されていない大部分の知識が構造化され、より豊かな情報サービスが提供される第一歩となる。

## 1.3 論文の構成

本論文は、全5章で構成されている。第1章では本研究の背景と目的について述べた。第2章では関連する研究や基盤となる技術について述べ、本研究の立ち位置を示す。第3章では提案するナレッジグラフ抽出手法を説明し、第4章では実験並びに評価と考察を行う。第5章では本論文を通してのまとめ、及び今後の展開について述べる。

## 第2章

# 関連研究・基盤技術

### 2.1 関連研究

大規模言語モデルとナレッジグラフがテーマである関連研究を3つ紹介する。1つ目は、グラフ構築に関わる基礎知識を大規模言語モデルに事前学習させ、テキストからナレッジグラフを作成する手法の提案 [7] である。本研究では、グラフ構築に関わる基礎知識を大規模言語モデルに事前学習させることはしない。テキストと大規模言語モデルのみを用いてナレッジグラフを作成する手法を提案する。

2つ目は、大規模言語モデルにナレッジグラフを基にした知識を与え、大規模言語モデルのコンテンツ生成における欠点であるハルシネーションを軽減する手法の提案 [8] である。本研究とは、大規模言語モデルに対するナレッジグラフの扱いが異なる。既存のナレッジグラフを用いて大規模言語モデルを強化するのではなく、大規模言語モデルを用いて新たなナレッジグラフを作成する。作成したナレッジグラフの活用先の1つとして、大規模言語モデルの強化は考えられる。

3つ目は、大規模言語モデルを活用して、既存のナレッジグラフに欠けている情報を補完する手法の提案 [9] である。本研究では、既存のナレッジグラフは利用せず、テキストデータのみを用意する。そして、そのテキストデータから新規のナレッジグラフを構築する。

### 2.2 基盤技術

#### 2.2.1 ナレッジグラフ

ナレッジグラフは、情報や知識を体系的に整理し、相互関係をグラフ構造で表現するための基盤技術である。この技術は、膨大なデータや情報の中から意味的なつながりを見つけ出し、有益な知識を抽出するのに役立つ。ナレッジグラフは、様々なドメインにわたる

情報をノードとエッジからなるグラフで表現する。各ノードは個別のエンティティや概念を示し、エッジはエンティティや概念の関係性を示す。この仕組みによって、異なる情報要素が複雑なつながりを持ち、知識の深層を探索することを可能にする。ナレッジグラフの利点として、情報を階層構造や網羅的な関連性で表現することが挙げられる。これにより、特定のトピックに関する包括的な理解が得られ、情報検索や質問応答システム、推論などの応用が容易になる。また、異なるデータソースからの情報を統合し、複雑な問題に対処する際にも有効である。ナレッジグラフは自然言語処理や機械学習の分野においても広く活用されており、質問応答、意味理解、文脈把握などの課題において高度な性能を発揮している。また、セマンティックウェブや推論エンジンと組み合わせることで、知識の自動化や効率的な情報処理が可能となる。総じて、ナレッジグラフは複雑な知識を体系的かつ可視化された形で管理し、それに基づいて効果的な意思決定や問題解決を支援するための重要な基盤技術である。

## 2.2.2 大規模言語モデル

大規模言語モデルは、自然言語処理の分野において革新的な進展をもたらしている基盤技術である。これらは、数十億以上のパラメータを持ち、膨大なデータセットから学習された人工知能のモデルである。代表例としては、OpenAI が開発した GPT シリーズが挙げられる。大規模言語モデルは、言語の理解、生成、翻訳などのタスクにおいて高度な性能を発揮する。さらに、トークンの関係性や文脈を理解し、自然言語のパターンや構造を捉えることも可能である。これにより、人間のような文章生成や多様な自然言語処理タスクにおいて、先進的な結果を実現している。この技術が注目を集める理由の1つは、事前学習と転移学習のアプローチにある。モデルは膨大なデータセットで先行学習を行い、一般的な言語理解を有する。その後、特定のタスクにおいては、限られたデータで微調整を行い、高い性能を獲得する。このようにして、コンピュータプログラムと自然言語とのインタラクションを向上させ、検索エンジン、対話型 AI アシスタント、文章生成、要約、翻訳、質問応答などの様々な応用分野において革新的な成果をもたらしている。



## 第 3 章

# 大規模言語モデルを用いたナレッジグラフ抽出手法

### 3.1 概要

大規模言語モデルを活用して、テキストデータからナレッジグラフを抽出する手法を説明する。大規模言語モデルは OpenAI の gpt-3.5-turbo-1106、テキストデータはロイターニュースのデータセットの 1 つである Reuters-21578<sup>\*1</sup>を用いる。テキストから抽出するエンティティは、Organization、Person、Location、Commodity の 4 属性とした。ニュース記事の話題の中心は組織と人物であり、それらを説明付ける基礎要素に地域と商品があるからである。リレーションは、WorksFor、CompetesWith、CollaboratesWith、IsPartOf、OperatesIn、Produces、Consumes の 7 種とした。組織や人物の間の関係としては、雇用関係、競合関係、協力関係、集約関係が基本であり、地域や商品と繋ぐための関係を加えた。出力するナレッジグラフは、その RDF トリプルを Turtle 形式で表現する。一連のプロセスは、準備、エンティティの抽出、エンティティの属性分類、リレーションの抽出、RDF トリプルの記述、以上 5 つで構成され、各プロセスに分けて説明していく。

### 3.2 提案手法

#### 3.2.1 準備

グラフ構造化知識の事前学習なし且つ自然言語処理を用いてナレッジグラフを抽出するパイプライン [10] を参考に、大規模言語モデルを用いてテキストからナレッジグラフを抽

---

<sup>\*1</sup> <https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>

出する。一連の流れは準備、エンティティの抽出、エンティティの属性分類、リレーションの抽出、RDF トリプルの記述の 5 段階で構成される。まずは、準備段階について説明していく。主な内容は、テキストデータの準備と大規模言語モデルとの会話環境の構築である。具体的な実装に関しては、gpt-3.5-turbo-1106 を活用するべく、大規模言語モデルの機能拡張を効率的に実装するためのライブラリである LangChain<sup>\*2</sup>を用いた。記述言語は Python である。

プログラムの準備段階 (リスト 3.1) について説明する。因みに、プログラムの全貌は付録 A で確認できる。まずは、チャット、過去のチャットを保存するメモリ、そしてその 2 つの組み合わせであるチェーンを初期化する。ここで、gpt のモデルを gpt-3.5-turbo-1106 とし、gpt の回答に含まれる確率変数を制御する temperature と seed を 0 とし、最後に OpenAI の API キーを与える。temperature と seed を 0 としているが、gpt-3.5-turbo-1106 は同じインプットに対して一語一句同じアプットを得ることは不可能な設計となっている。従って、回答結果には多少のブレが生じることは述べておくが、今回の趣旨において問題にはならない。因みに、当初利用していた gpt-3.5 モデルは、temperature か seed を 0 とすれば、同じインプットに対して一語一句同じアウトプットを得られていたが、現在は利用不可である。

リスト 3.1: プログラム準備段階

```
21 #Initialize chat
22 chat = ChatOpenAI(model_name="gpt-3.5-turbo-1106",
23                   temperature=0,
24                   seed=0,
25                   openai_api_key="your_api_key")
26
27 #Initialize memory
28 memory = ConversationBufferMemory()
29
30 #Initialize chain
31 conversation = ConversationChain(
32     llm=chat,
33     memory=memory
34 )
```

### 3.2.2 エンティティの抽出

LangChain を用いて、チャットとメモリ、そしてチェーンを準備したので、チャットに与える命令文であるプロンプトを記述していく。プロンプトとそれに対するレスポンス

---

<sup>\*2</sup> <https://www.langchain.com>

は表 3.1 にまとめる。まずは、次に与えるテキストから固有名詞を抽出してもらうことを指示する。ここで、ナレッジグラフで抽出すべきものを表すエンティティではなく、固有名詞としたのには理由がある。エンティティの抽出を命令すると、大量の普通名詞やテキストデータとしたニュース記事には数多く登場する数字まで羅列される。今回はテキストから、組織、人物、地域、商品の 4 属性のみを抽出するため、普通名詞や数字は必要なく、必要となる単語をより正確に表す固有名詞の抽出を指示した。

次に与えるプロンプトは、テキストデータであり、プログラム内の変数に格納したものを呼び出してチャットに与えている。具体例として、データセットからランダムに選んだテキストを用いたものを示す。この記事は、米国自動車市場について記述されたものである。レスポンスから、テキストに含まれる固有名詞がエンティティとして正しく抽出されていることが分かる。

### 3.2.3 エンティティの属性分類

エンティティを抽出したので、その属性进行分类する。属性は概要で説明した通り、Organization、Person、Location、Commodity の 4 つである。プロンプトとそれに対するレスポンスは表 3.2 にまとめてある。まず、前のステップで抽出したエンティティを、組織属性に当てはまるものに制限する。そして、同様の制限を、人物属性、地域属性、商品属性でも行う。結果、4 つの命令に対して期待されるレスポンスが得られ、抽出したエンティティから、必要となる属性に該当する単語を属性別に取得することができた。

表 3.1: 大規模言語モデルとの会話：エンティティの抽出

入力 1:	Can you extract the proper nouns from the text I give next?
GPT:	Sure, I can do that. Please provide the text you would like me to analyze.
入力 2:	テキストデータ
GPT:	The proper nouns in the text you provided are: Geneva Motor Show, U.S., Ford Motor Co, General Motors Corp, Europe, North American, Chrysler Corp, Britain, West Germany, Opel/Vauxhall, Spain, James Fry, North American, Reuters, Walter Hayes, Ford Europe, Robert Lutz, Chrysler, European.

表 3.2: 大規模言語モデルとの会話：エンティティの属性分類

入力 3:	Can you restrict the previous result to those belonging to “Organization”?
GPT:	The proper nouns belonging to “Organization” in the text you provided are: Geneva Motor Show, Ford Motor Co, General Motors Corp, Chrysler Corp, Opel/Vauxhall, Ford Europe, Reuters.
入力 4:	Can you restrict the previous result to those belonging to “Person”?
GPT:	The proper nouns belonging to “Person” in the text you provided are: James Fry, Walter Hayes, Robert Lutz.
入力 5:	Can you restrict the previous result to those belonging to “Location”?
GPT:	The proper nouns belonging to “Location” in the text you provided are: Europe, Britain, West Germany, Spain.
入力 6:	Can you restrict the previous result to those belonging to “Commodity”?
GPT:	The proper nouns belonging to “Commodity” in the text you provided are: North American.

### 3.2.4 リレーションの抽出

エンティティとその属性の抽出が完了したので、エンティティ間のリレーションを抽出する。リレーションは概要で説明した通り、WorksFor、CompetesWith、CollaboratesWith、IsPartOf、OperatesIn、Produces、Consumes の 7 種である。また、命令をする際に、各リレーションの複合動詞に対する主語と目的語の属性を制限している。これは、ナレッジグラフを構築する際に行われる、基本的な RDF トリプルの制限に値する。プロンプトとそれに対するレスポンスは表 3.3, 3.4 にまとめてある。まずは、人物属性を主語にもち、組織属性を目的語にもつ、WorksFor 関係を抽出する。そして、属性とリレーションの組み合わせ他 7 つに対しても同様の操作をする。このようにして、エンティティ間のリレーションを特定した。

### 3.2.5 RDF トリプルの記述

テキストからエンティティとそれらのリレーションを抽出できたので、RDF トリプルを Turtle 形式で記述する。記述形式を一意に定めるため、“@prefix ex:” から記述をさせるようにした。プロンプトとそれに対するレスポンスは表 3.5 にまとめてあり、抽出したエンティティのリレーションを表す RDF トリプルを Turtle 形式で得ることができた。

表 3.3: 大規模言語モデルとの会話：リレーションの抽出

入力 7 :	Can you describe the relationships that apply “Person WorksFor Organization”?
GPT :	<p>The relationships that apply “Person WorksFor Organization” in the text you provided are:</p> <ul style="list-style-type: none"> <li>- James Fry WorksFor GM</li> <li>- Walter Hayes WorksFor Ford Europe</li> <li>- Robert Lutz WorksFor Chrysler</li> </ul>
入力 8 :	Can you describe the relationships that apply “Organization CompetesWith Organization”?
GPT :	<p>The relationships that apply “Organization CompetesWith Organization” in the text you provided are:</p> <ul style="list-style-type: none"> <li>- Ford Motor Co CompetesWith General Motors Corp</li> <li>- Ford Motor Co CompetesWith Opel/Vauxhall</li> <li>- General Motors Corp CompetesWith Ford Motor Co</li> <li>- General Motors Corp CompetesWith Opel/Vauxhall</li> <li>- Chrysler Corp CompetesWith Ford Motor Co</li> <li>- Chrysler Corp CompetesWith General Motors Corp</li> </ul>
入力 9 :	Can you describe the relationships that apply “Organization CollaboratesWith Organization”?
GPT :	<p>The relationships that apply “Organization CollaboratesWith Organization” in the text you provided are:</p> <ul style="list-style-type: none"> <li>- Ford Motor Co CollaboratesWith Ford Europe</li> </ul>
入力 10 :	Can you describe the relationships that apply “Person CollaboratesWith Person”?
GPT :	I’m sorry, I cannot find any relationships that apply “Person CollaboratesWith Person” in the text you provided.

表 3.4: 大規模言語モデルとの会話：リレーションの抽出 (続き)

---

入力 11 :	Can you describe the relationships that apply “Organization IsPartOf Organization”?
GPT :	The relationships that apply “Organization IsPartOf Organization” in the text you provided are: - Ford Europe IsPartOf Ford Motor Co

---

入力 12 :	Can you describe the relationships that apply “Organization OperatesIn Location”?
GPT :	The relationships that apply “Organization OperatesIn Location” in the text you provided are: - Ford Motor Co OperatesIn Europe - General Motors Corp OperatesIn Europe - Chrysler Corp OperatesIn Europe - Ford Europe OperatesIn Britain - Ford Europe OperatesIn West Germany

---

入力 13 :	Can you describe the relationships that apply “Organization Produces Commodity”?
GPT :	The relationships that apply “Organization Produces Commodity” in the text you provided are: - Ford Motor Co Produces North American - General Motors Corp Produces North American - Chrysler Corp Produces North American

---

入力 14 :	Can you describe the relationships that apply “Organization Consumes Commodity”?
GPT :	I’m sorry, I cannot find any relationships that apply “Organization Consumes Commodity” in the text you provided.

表 3.5: 大規模言語モデルとの会話：RDF トリプルの記述

---

入力 15 :	Can you describe all the relationships in Turtle format triples. Please output in the format starting from “@prefix ex”?
GPT :	<p>@prefix ex: &lt;http://example.org/&gt; .</p> <p>ex:JamesFry ex:WorksFor ex:GM .</p> <p>ex:WalterHayes ex:WorksFor ex:FordEurope .</p> <p>ex:RobertLutz ex:WorksFor ex:Chrysler .</p> <p>ex:FordMotorCo ex:CompetesWith ex:GeneralMotorsCorp .</p> <p>ex:FordMotorCo ex:CompetesWith ex:OpelVauxhall .</p> <p>ex:GeneralMotorsCorp ex:CompetesWith ex:FordMotorCo .</p> <p>ex:GeneralMotorsCorp ex:CompetesWith ex:OpelVauxhall .</p> <p>ex:ChryslerCorp ex:CompetesWith ex:FordMotorCo .</p> <p>ex:ChryslerCorp ex:CompetesWith ex:GeneralMotorsCorp .</p> <p>ex:FordMotorCo ex:CollaborateWith ex:FordEurope .</p> <p>ex:FordEurope ex:IsPartOf ex:FordMotorCo .</p> <p>ex:FordMotorCo ex:OperatesIn ex:Europe .</p> <p>ex:GeneralMotorsCorp ex:OperatesIn ex:Europe .</p> <p>ex:ChryslerCorp ex:OperatesIn ex:Europe .</p> <p>ex:FordEurope ex:OperatesIn ex:Britain .</p> <p>ex:FordEurope ex:OperatesIn ex:WestGermany .</p> <p>ex:FordMotorCo ex:Produces ex:NorthAmerican .</p> <p>ex:GeneralMotorsCorp ex:Produces ex:NorthAmerican .</p> <p>ex:ChryslerCorp ex:Produces ex:NorthAmerican .</p>



## 第 4 章

# 実験・評価・考察

### 4.1 実験

データセット<sup>\*1</sup>からランダムに選択した 10 個のテキストから、2 つの異なる手法でナレッジグラフを抽出して比較する実験を行う。一方は、自然言語処理を用いた従来手法 [10] であり、もう一方は、大規模言語モデルを用いた提案手法である。実験の目的は、テキストデータからナレッジグラフを抽出するにおいて、どちらの手法がより適しているかを判断し、提案手法の有用性を検証することである。その方法として、従来手法と提案手法で抽出したナレッジグラフに対して、その情報量と正答率を数値化して比較する。情報量はエンティティを表すノードとリレーションを表すエッジの数で判断、正答率はトリプルが示す内容がテキストから読み取れるかで判断する。実験の結果は表 4.1, 4.2 の通りである。

### 4.2 評価

表 4.1, 4.2 に示した実験結果を元に、従来手法と提案手法を比較する。まずは、エンティティの数を表すノード数についてであるが、すべての記事で提案手法が従来手法に比べて、同じかそれ以上となっており、提案手法が従来手法に劣っているテキストはない。そして、ノード数の平均は、4.9 から 10.9 へと約 2.2 倍になり、情報量が大きく増えている。つぎに、関係性を表すエッジ数についてであるが、こちらもすべての記事で提案手法が従来手法に比べて、同じかそれ以上となっている。エッジ数の平均は、3.9 から 9.9 へと約 2.5 倍となっており、情報量の伸びはノード数よりも大きい。

つづけて正答率について評価する。正答率は、従来手法で 34/39 個であり約 87.1%、提案手法で 98/99 個で約 98.9% となっている。正答率でも提案手法は従来手法を上回り、

---

<sup>\*1</sup> <https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>

大規模言語モデルの自然言語理解能力が正しく発揮されていることがうかがえる。以上より、テキストデータからナレッジグラフを抽出する手法として、従来手法よりも提案手法の有用性が高いと言える。

表 4.1: ノード数

	従来手法	提案手法
テキスト 1	5	14
テキスト 2	4	12
テキスト 3	4	10
テキスト 4	10	14
テキスト 5	4	8
テキスト 6	3	3
テキスト 7	3	4
テキスト 8	5	12
テキスト 9	6	26
テキスト 10	5	6
平均	4.9	10.9

表 4.2: 正答数/エッジ数

	従来手法	提案手法
テキスト 1	4/4	19/19
テキスト 2	4/4	10/10
テキスト 3	3/3	9/9
テキスト 4	9/9	11/11
テキスト 5	2/3	6/6
テキスト 6	1/2	3/3
テキスト 7	2/2	2/2
テキスト 8	3/4	11/12
テキスト 9	2/4	21/21
テキスト 10	4/4	6/6
平均	3.4/3.9	9.8/9.9

## 4.3 考察

### 4.3.1 提案手法の優位点

実験の結果を考察する。提案手法が従来手法に、情報量と正確性の両方で優位となった理由は、文脈把握能力並びに知識抽出精度の差だと考える。それらを証明する例を、ナレッジグラフやテキスト内容を必要に応じて含め、挙げていく。

#### 記述されない競合関係

文章では明記されていない競合関係についてである。“企業 A と企業 B は競合である”という文章は、その事実が自明である際は記述されない。例えば、GM、Ford、Chrysler の自動車企業についての記事から、各企業が互いに Compete 関係であることは、従来手法 (図 4.1) では抽出できていないが、提案手法 (図 4.2) では抽出できている。同様に、Quebecor Inc、New York Times、Daily Mirror の出版社についての記事から、各企業が互いに Compete 関係であることは、従来手法 (図 4.3) では抽出できていないが、提案手法 (図 4.4) では抽出できている。

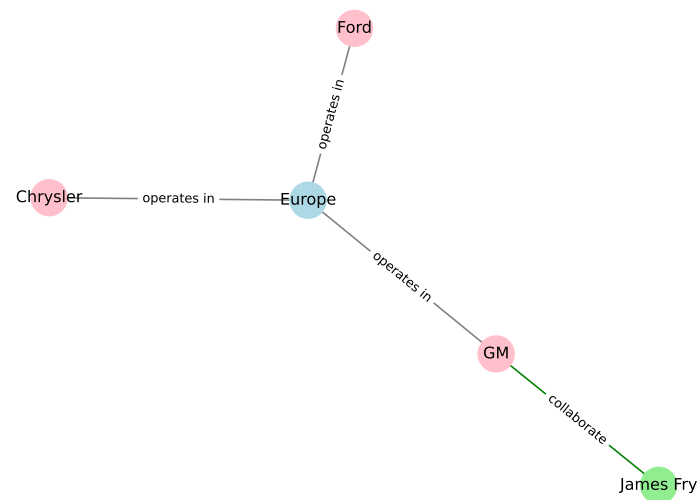


図 4.1: 従来手法 (テキスト 1)  
米国自動車メーカーに関するナレッジグラフ

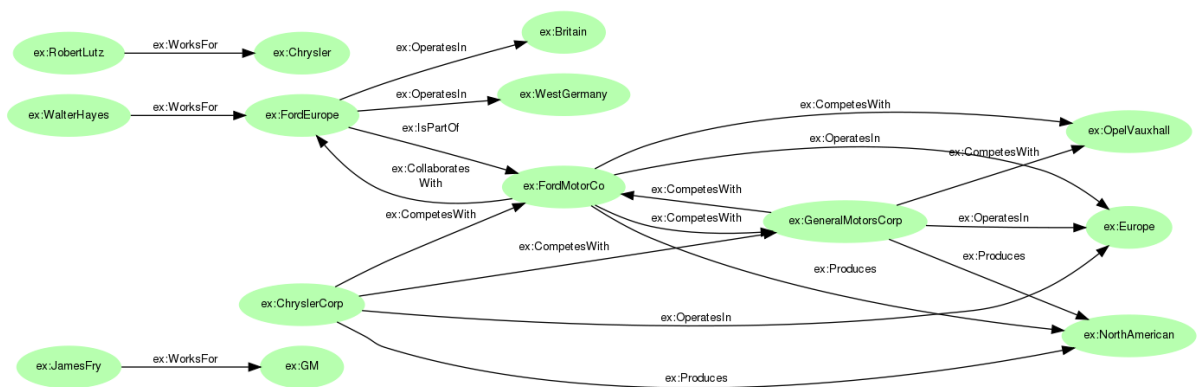


図 4.2: 提案手法 (テキスト 1)  
米国自動車メーカーに関するナレッジグラフ

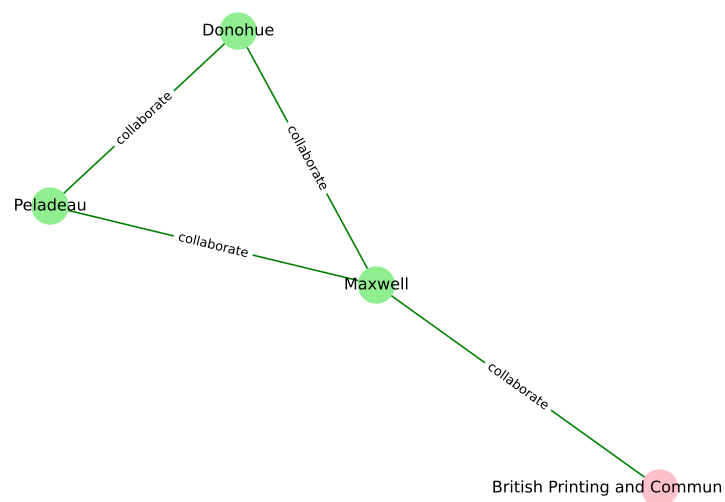


図 4.3: 従来手法 (テキスト 2)  
加国出版社 Quebecor に関するナレッジグラフ



図 4.4: 提案手法 (テキスト 2)  
加国出版社 Quebecor に関するナレッジグラフ

## 生産関係

商品の生産関係の抽出精度についてである。テキスト 1 の “Chrysler はアメリカ製自動車の輸出を開始する” や “GM はヨーロッパにおける北米製車の販売台数を 5 倍に増加”、テキスト 2 の “カナダ最大の出版印刷会社のひとつである Quebecor Inc”、テキスト 3 の “今週のジュネーブモーターショーで展示されるのは、クライスラーの Lancer ES、Plymouth Voyager、Dodge Daytona Shelby Z である” といった内容の文章が例に挙げられる。従来手法 (図 4.1,4.3,4.5) ではこれらから、企業と商品の Produce 関係を 1 つも抽出できていないのに対し、提案手法 (図 4.2,4.4,4.6) ではすべての Produce 関係を抽出できている。

### テキスト 1 抜粋

Now Chrysler Corp <C> is saying it will begin exporting American-made vehicles before the end of this year to Europe, a market it left in 1978 when it was near bankruptcy.

James Fry of GM overseas distribution system said GM hoped for a five-fold rise this year in sales of North American-made vehicles in Europe, selling between 7,000 and 8,000 (North American-made) units in Europe for the year to August 1987.

### テキスト 2 抜粋

Quebecor Inc, one of Canada's largest publishing and printing companies, is likely to launch a new daily newspaper in Montreal, probably this fall, president Pierre Peladeau told Reuters in an interview.

### テキスト 3 抜粋

The vehicles, on show at this week's Geneva motor show, include Chrysler LeBaron Turbo coupe and convertible, the Lancer ES four-door hatchback, the Dodge Shadow ES compact, the Plymouth Voyager minivan and the Dodge Daytona Shelby Z.

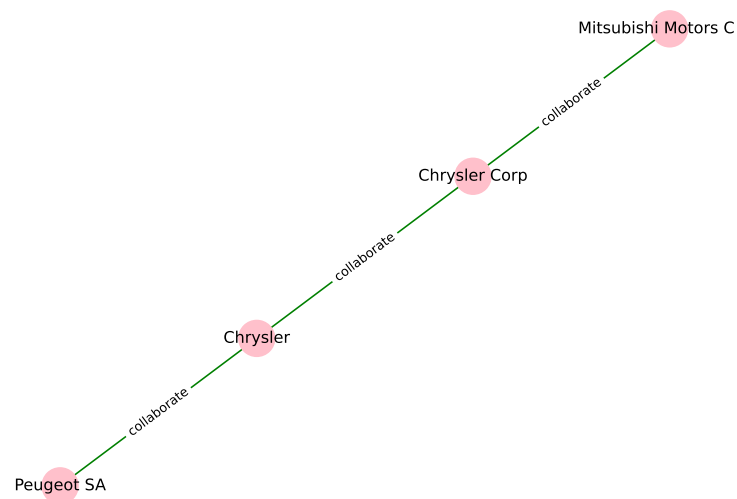


図 4.5: 従来手法 (テキスト 3)  
自動車メーカー Chrysler に関するナレッジグラフ

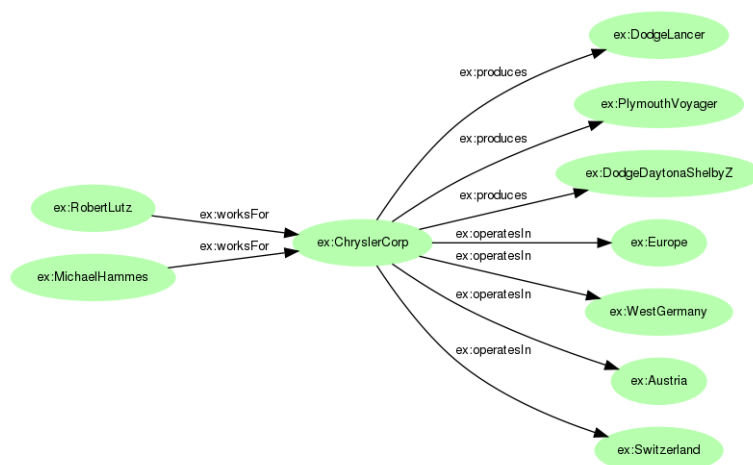


図 4.6: 提案手法 (テキスト 3)  
自動車メーカー Chrysler に関するナレッジグラフ

### 複数の意味をもつ関係

テキスト 4 の“錫生産国連合 (ATPC) 非加盟国のブラジルと中国は ATPC に協力し、輸出制限を約束した”という内容の文章から、ATPC とブラジル並びに中国は、Compete 関係でありながら Collaborate 関係であるということを、提案手法 (図 4.8) では正しく抽出できている。これは、提案手法の文脈抽出能力の高さを表す良い例の 1 つである。ちなみに従来手法 (図 4.7) では、この内容の文章から何も情報を抽出していない。

テキスト 4 抜粋

Non-members Brazil and China have pledged to cooperate with the ATPC and limit their exports to 21,000 and 7,000 tonnes respectively during the quota period.

### 取材関係と競合関係

従来手法 (図 4.7,4.9) では、テキスト 4 での“Dam Teutong が Reuters に語った”や“Jaimie Villalobos が Reuters に語った”、テキスト 5 での“James Fry が Reuters に語った”などを意味する文章から、各人物と Reuters が Compete 関係であるとしている。彼らは記者の取材に答えたのであり、争っていない。これは従来手法が文脈を正しく理解できていないことを意味する。ちなみに、提案手法 (図 4.8,4.10) ではこの文脈は抽出されておらず、語っただけでは Collaborate 関係にも値しないと判断したと予測される。

テキスト 4 抜粋

Chief Inspector of Mines Redzuan Sumun told Reuters that excess output of some 3,000 tonnes after exports of 28,500 in the one-year quota period would be kept in the national stock.

A spokesman for the state-owned tin mining company PT Tambang Timah told Reuters it would be easy for Indonesia to stick to an export quota of 24,516 tonnes because this was close to exports in calendar 1986 of 24,636.

テキスト 5 抜粋

“We would like to sell between 7,000 and 8,000 units in Europe for the year to August 1987,” he told Reuters later.



図 4.7: 従来手法 (テキスト 4)  
錫生産国連合 ATPC に関するナレッジグラフ

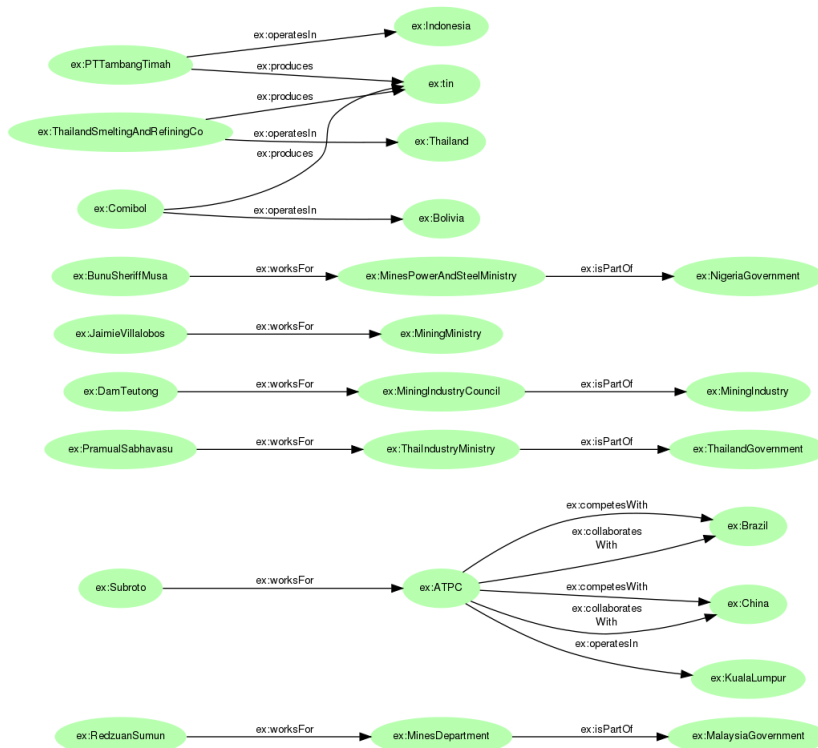


図 4.8: 提案手法 (テキスト 4)  
錫生産国連合 ATPC に関するナレッジグラフ



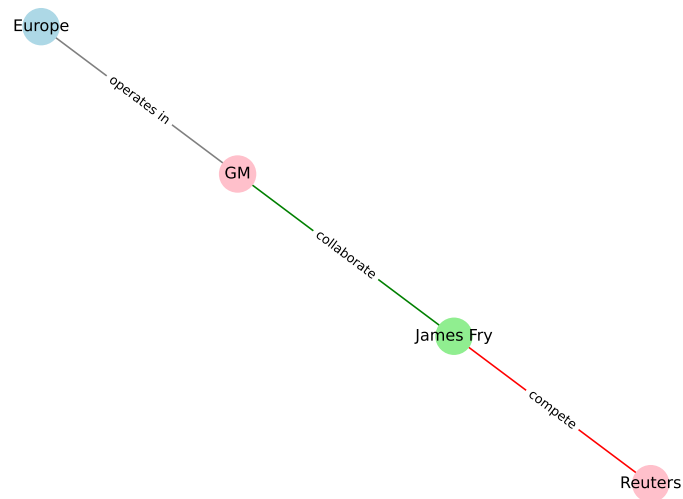


図 4.9: 従来手法 (テキスト 5)  
自動車メーカー GM に関するナレッジグラフ

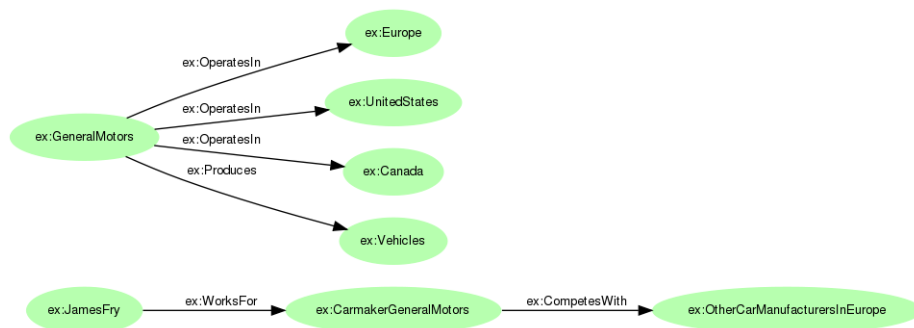


図 4.10: 提案手法 (テキスト 5)  
自動車メーカー GM に関するナレッジグラフ

## 並列関係

テキスト 6 の “Outokumpu の部門は、Metallverken と Wirsbo の銅生産工場を組み込む” という内容の文章で、従来手法 (図 4.11) は、Metallverken Collaborate Wirsbo、Wirsbo collaborate Outokumpu としており、並列関係を正しく理解できていない。提案手法 (図 4.12) では、Outokumpu Collaborate Metallverken、Outokumpu Collaborate Wirsbo となっており、正しく知識を抽出できている。

テキスト 6 抜粋

The new Outokumpu division, called Copper Products Industry, is to incorporate Outokumpu's copper production plants, including its two U.S. Subsidiaries <Nippert Co> and <Valleycast Inc>, as well as Metallverken and part of Wirsbo.

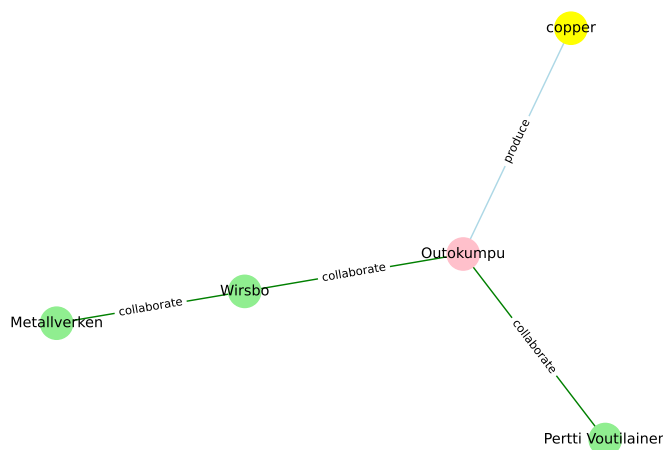


図 4.11: 従来手法 (テキスト 6)

鉱業企業 Outokumpu に関するナレッジグラフ



図 4.12: 提案手法 (テキスト 6)

鉱業企業 Outokumpu に関するナレッジグラフ

## 分類難易度の高い関係

テキスト7の“米国農務省の発表によると、商品信用公社は、モロッコ向け米国産小麦の販売に対し、4,500万ドルを追加許可した”という文章から、商品信用公社が小麦を Consume するという関係を従来手法 (図 4.13) は抽出している。これは人間でもある程度理解の難しい文章ではあるかもしれないが、少なくとも商品信用公社が小麦を Produce しているとは言えるが、Consume 関係は見いだせない。このような複雑な文脈の文章を、従来手法の自然言語処理は苦手とする。ちなみに、提案手法 (図 4.14) では商品信用公社はモロッコで OperateIn しているという情報を抽出している。これは OperateIn の定義を厳密にすれば、誤りであるかもしれないが、商品信用公社がモロッコに対して影響力を持っていることは明らかであり、その意味で OperateIn は正しい知識として機能する。

テキスト7抜粋

The Commodity Credit Corporation has approved an additional 45.0 mln dlrs under its interemdiat export credit guarantee program, GSM-103, for sales of U.S. wheat to Morocco, the U.S. Agriculture Department said.

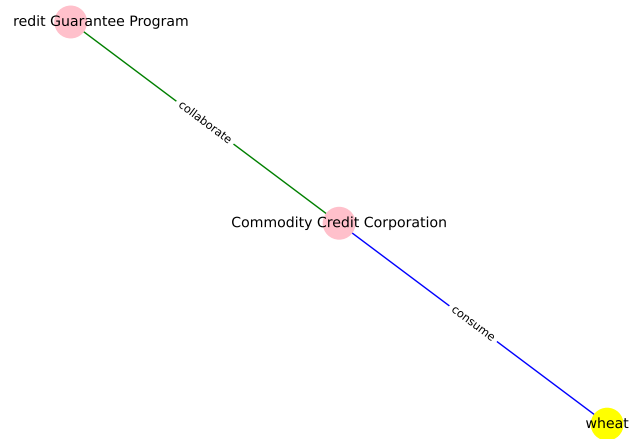


図 4.13: 従来手法 (テキスト 7)

米国商品信用公社に関するナレッジグラフ

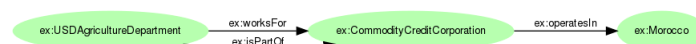


図 4.14: 提案手法 (テキスト 7)

米国商品信用公社に関するナレッジグラフ

### 4.3.2 提案手法の欠点

大規模言語モデルを用いた提案手法の文脈把握能力と知識抽出精度の高さを示してきたが、欠点も2点挙げられる。

#### 単語統一

1つ目は、単語の統一であり、テキスト1では“Ford Europe”と“Ford Motor Co”の2つのエンティティが登場し、“Ford”として統一はされていない(図4.2)。従来手法(図4.1)では相関係数の高い単語らは、そのうち最も簡単な表現の単語として統一する処理がなされている。プロンプトを加えることで、同様の処理の実現を試みたが、上手く実装することはできなかった。その代替案として、従来手法にはないIsPartOf関係を追加して抽出した。これにより、テキスト1の“Ford Europe”と“Ford Motor Co”は相関の高い関係であることがナレッジグラフ(図4.2)から読み取れるようになり、問題は解決された。また、この実装の良い副作用があり、テキスト2の“Donohue Incを買収する入札を落札した Quebecor Inc”、テキスト4での“タイの Pramual Sabhavasus 工業大臣”や“Bunu Sheriff Musa 鉱山電気鉄鋼相はラゴスで、ナイジェリアの状況について述べた”などの内容の文章から、Donohue IncとQuebecor Inc、工業省とタイ政府、鉱山電気鉄鋼省とナイジェリア政府がisPartOf関係であることが追加抽出され、ナレッジグラフの情報量が増えると共に、全体知識が理解しやすいものとなった(図4.4,4.8)。

#### テキスト2 抜粋

The company, which last week won a bid to buy the Quebec government's 55 pct interest in pulp and paper company <Donohue Inc>, will also likely go ahead with plans to build a new paper mill in Matane, Quebec, Peladeau said.

#### テキスト4 抜粋

In Bangkok, Thai Industry Minister Pramual Sabhavasus said Thailand would keep to its 19,000 tonne quota and added this would not cause the local industry hardship at current prices.

Mines, Power and Steel Minister Bunu Sheriff Musa said in Lagos that Nigeria would have no difficulty keeping within its ATPC quota of 1,461 tonnes because its metals output had declined due to poor demand and high production costs.

## 属性制限

2つ目は、エンティティの属性制限である。各リレーションの主語と目的語になるエンティティの属性を制限するように、プロンプトを記述してあるのだが、低い確率でその制限を無視した回答が得られた。テキスト 6 の “Outokumpu Oy はスウェーデンの銅メーカー 2 社を買収し、銅加工部門に統合した” という内容の文章から、Outokumpu Oy と銅加工部門に Produce 関係と IsPartOf 関係を結んでいる (図 4.12)。IsPartOf 関係は問題ないのだが、Produce 関係の目的語は Commodity 属性に制限しており、銅加工部門はあてはまらない。Produce という文字で見れば正しい関係であるが、ここでは商品以外の目的語は期待しておらず、制御が正しく機能していない。このような事象はプロンプトエンジニアリング以外では発生せず、大規模言語モデルを用いて要素制限を実装する際の弱点と言える。

テキスト 6 抜粋

Outokumpu acquired two Swedish copper manufacturers in January 1986, <Metallverken Ab> and <Wirsbo Bruks Ab>, that were merged into its copper processing division.



## 第5章

# おわりに

本論文では、大規模言語モデルを用いて、テキストデータからナレッジグラフを抽出する手法を提案した。実装は、従来の自然言語処理を用いた手法における各プロセスを参考にしたプロンプトを、大規模言語モデルに逐次的に与えて実現した。そして、従来手法との比較実験により、提案手法が情報量と正確性の両面で優位であることが示された。それは、文脈把握能力と知識抽出精度の高さが要因であると考察でき、それらを証明する例を示した。

今後の展開について述べる。まずは、最適プロンプトの更なる模索である。欠点としてあげた、単語統一と属性制限の問題に焦点を当てたい。信頼できるナレッジグラフ作成のため、特に後者は発生率は低いとしても優先的に解決される必要がある。つぎに、大規模言語モデルの強化による恩恵の享受である。大規模言語モデルはアップデートが頻繁に行われ、学習量の増した強化モデルが随時提供される。それにより、文脈把握能力と知識抽出精度が更に増し、より実用的なナレッジグラフが構築できるようになるであろう。最後に、他テキストデータのナレッジグラフ化である。未だ構造化されていない知識のテキストデータは山ほどある。それらの構造化することにより、提案手法のブラッシュアップとともに、知識情報サービスの更なる発展が期待できる。





# 謝辞

本研究の遂行及び本論文の執筆にあたり、終始熱心なご指導をいただいた桑原和宏教授に深く感謝申し上げます。また、研究室で多くの時間をともにし、多くの助言と激励をいただいた Didier Gohourou 氏にも心より感謝申し上げます。最後に、日々大きな活力をいただいた親愛なる友人諸君と家族、そして快適な大学運営に携わってくださった全ての方々に感謝の意を表します。



# 参考文献

- [1] Ellen Riloff, et al. Automatically constructing a dictionary for information extraction tasks. In *AAAI*, Vol. 1, pp. 2–1, 1993.
- [2] Soufyane Ayanouz, Boudhir Anouar Abdelhakim, and Mohammed Benhmed. A smart chatbot architecture based nlp and machine learning for health care assistance. In *Proceedings of the 3rd international conference on networking, information systems & security*, pp. 1–6, 2020.
- [3] Zhichen Zeng, Ruike Zhu, Yinglong Xia, Hanqing Zeng, and Hanghang Tong. Generative graph dictionary learning. In *International Conference on Machine Learning*, pp. 40749–40769. PMLR, 2023.
- [4] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, 2019.
- [5] Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (llm)? a.k.a. will llms replace knowledge graphs?, 2023.
- [6] Qing Huang, Zhenyu Wan, Zhenchang Xing, Changjing Wang, Jieshan Chen, Xiwei Xu, and Qinghua Lu. Let’s chat to find the apis: Connecting human, llm and knowledge graph through ai chain. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 471–483, 2023.
- [7] Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. *arXiv preprint arXiv:2310.13023*, 2023.
- [8] Ariana Martino, Michael Iannelli, and Coleen Truong. Knowledge injection to counter large language model (llm) hallucination. In Catia Pesquita, Hala Skaf-Molli, Vasilis Efthymiou, Sabrina Kirrane, Axel Ngonga, Diego Collarana, Renato Cerqueira, Mehwish Alam, Cassia Trojahn, and Sven Hertling, editors, *The Semantic Web: ESWC 2023 Satellite Events*, pp. 182–185, Cham, 2023.

Springer Nature Switzerland.

- [9] Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. Exploring large language models for knowledge graph completion, 2023.
- [10] Didier Gohourou and Kazuhiro Kuwabara. Knowledge graph extraction of business interactions from news text for business networking analysis. *Machine Learning and Knowledge Extraction*, Vol. 6, No. 1, pp. 126–142, 2024.

# 付録 A

## ソースコード

提案手法の実装内容である。

リスト A.1: ソースコード

```
1 from langchain.chains import ConversationChain
2 from langchain.memory import ConversationBufferMemory
3 from langchain.chat_models import ChatOpenAI
4
5 #Dictionary of article text data
6 article_dict = {
7     "article_1": ""text_data_article_1"",
8     "article_2": ""text_data_article_2"",
9     "article_3": ""text_data_article_3"",
10    "article_4": ""text_data_article_4"",
11    "article_5": ""text_data_article_5"",
12    "article_6": ""text_data_article_6"",
13    "article_7": ""text_data_article_7"",
14    "article_8": ""text_data_article_8"",
15    "article_9": ""text_data_article_9"",
16    "article_10": ""text_data_article_10""
17 }
18 #Run the following program in each article
19 for article in article_dict.values():
20
21     #Initialize chat
22     chat = ChatOpenAI(model_name="gpt-3.5-turbo-1106",
23                       temperature=0,
24                       seed=0,
25                       openai_api_key="your_api_key")
26
27     #Initialize memory
28     memory = ConversationBufferMemory()
29
30     #Initialize chain
```

```

31 conversation = ConversationChain(
32     llm=chat,
33     memory=memory
34 )
35
36 #Conversation 1
37 input_1 = "Can you extract the proper nouns from the text I
    give next?"
38 res = conversation.run(input_1)
39 #Command and response outputs
40 print("Input 1:" + input_1)
41 print(res)
42
43 #Conversation 2
44 #input_2 = article
45 res = conversation.run(article)
46 #Command and response outputs
47 print("Input 2:" + "The article text data")
48 print(res)
49
50 #Conversation 3 to 14
51 #Dictionary of Commands
52 input_dict = {
53     "input_3": "Can you restrict the previous result to those
        belonging to \"Organization?\"",
54     "input_4": "Can you restrict the previous result to those
        belonging to \"Person?\"",
55     "input_5": "Can you restrict the previous result to those
        belonging to \"Location?\"",
56     "input_6": "Can you restrict the previous result to those
        belonging to \"Commodity?\"",
57     "input_7": "Can you describe the relationships that apply \"
        Person WorksFor Organization?\"",
58     "input_8": "Can you describe the relationships that apply \"
        Organization CompetesWith Organization?\"",
59     "input_9": "Can you describe the relationships that apply \"
        Organization CollaborateWith Organization?\"",
60     "input_10": "Can you describe the relationships that apply \"
        Person CollaborateWith Person?\"",
61     "input_11": "Can you describe the relationships that apply \"
        Organization IsPartOf Organization?\"",
62     "input_12": "Can you describe the relationships that apply \"
        Organization OperatesIn Location?\"",
63     "input_13": "Can you describe the relationships that apply \"
        Organization Produces Commodity?\"",
64     "input_14": "Can you describe the relationships that apply \"
        Organization Consumes Commodity?\""
65 }

```

```

66     #Run each command
67     n = 3
68     for input in input_dict.values():
69         res = conversation.run(input)
70         #Command and response outputs
71         print("Input_␣"+str(n)+" :␣"+input)
72         print(res)
73         n += 1
74
75     #Conversation 15
76     input_15 = """Can you describe all the relationships in
77         Turtle format triples. Please output in the format
78         starting from "@prefix ex:"?"""
79     #Command and response outputs
80     res = conversation.run(input_15)
81     print("Input_␣15:␣"+input_15)
82     print(res)

```





## 付録 B

# ナレッジグラフ

実験の結果に含まれるが、本文に登場しなかったナレッジグラフを示す。

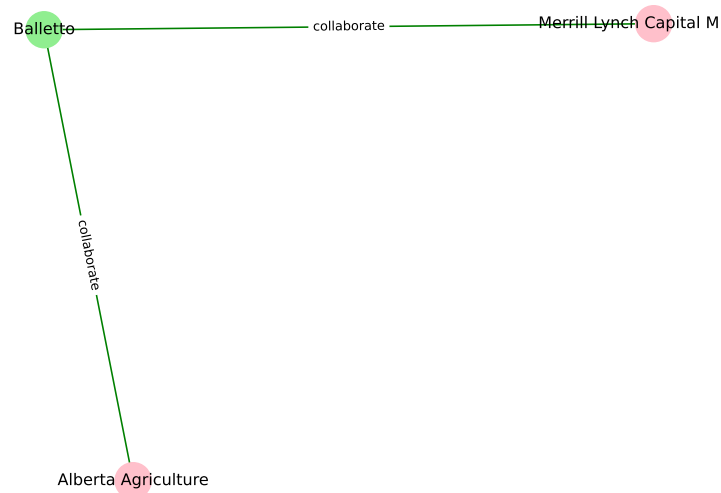


図 B.1: 従来手法 (テキスト 8)

アルバータ州農業に関するナレッジグラフ

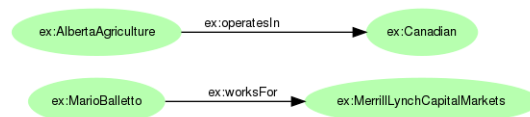


図 B.2: 提案手法 (テキスト 8)

アルバータ州農業に関するナレッジグラフ

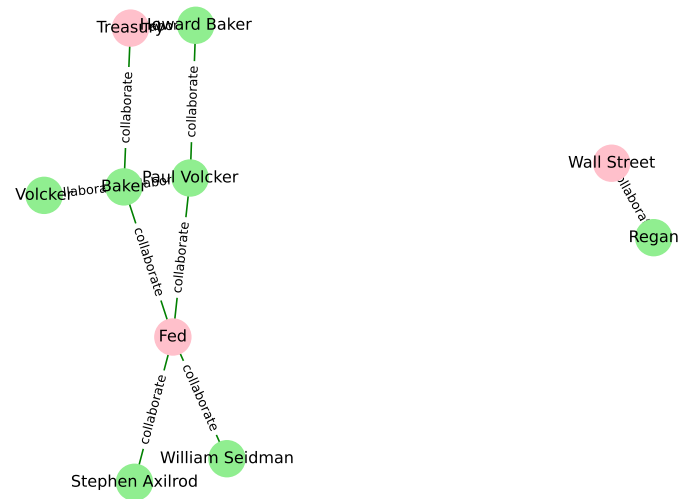


図 B.3: 従来手法 (テキスト 9)  
ホワイトハウス主席補佐官に関するナレッジグラフ

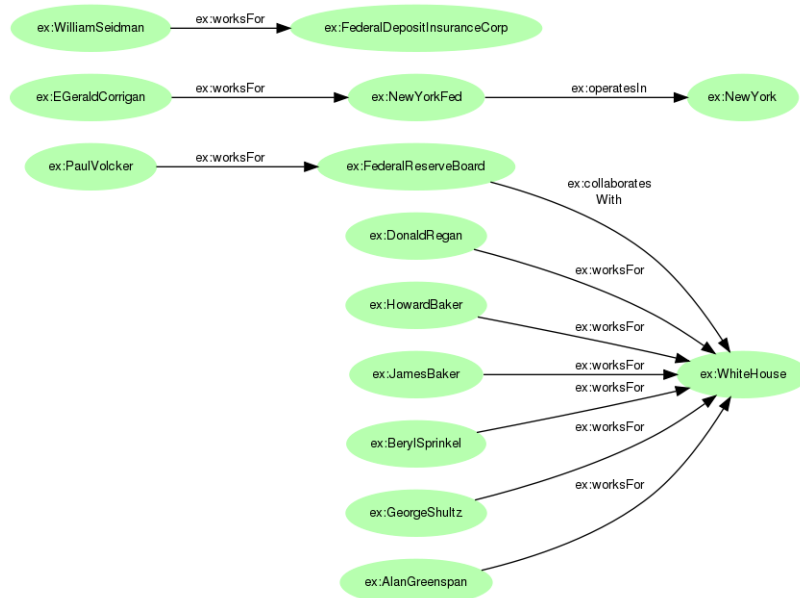


図 B.4: 提案手法 (テキスト 9)  
ホワイトハウス主席補佐官に関するナレッジグラフ

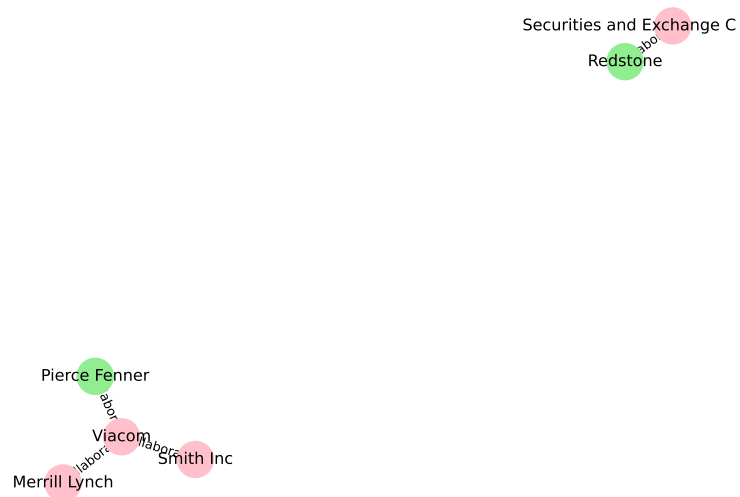


図 B.5: 従来手法 (テキスト 10)  
米国放送会社 Viacom に関するナレッジグラフ

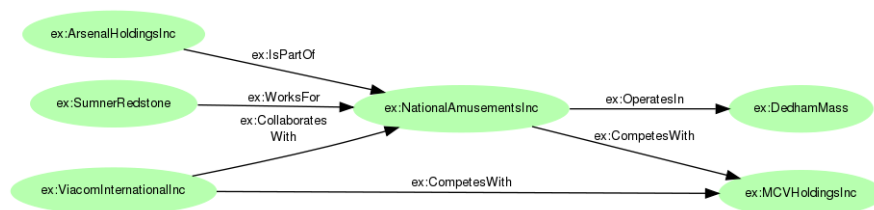


図 B.6: 提案手法 (テキスト 10)  
米国放送会社 Viacom に関するナレッジグラフ